SNP Identity Searching Across Ethnicities, Kinship, and Admixture

Authors: Brian S. Helfer & Darrell O. Ricke

ABSTRACT

Advances in high throughput sequencing (HTS) enables application of single nucleotide polymorphism (SNP) panels for identification and mixture analysis. Large reference sets of characterized individuals with documented relationships, ethnicities, and admixture are not yet available for characterizing the impacts of different ethnicities, kinships, and admixture on identification and mixture search results for relatives and unrelated individuals. Models for the expected results are presented with comparison results on two *in silico* datasets spanning four ethnicities, extended kinship relationships, and also admixture between the four ethnicities.

KEYWORDS

Forensic science, single nucleotide polymorphism, kinship, admixture, identification

Introduction

The United States recently expanded the Combined DNA Index System (CODIS) loci from 13 to 20 in 2016. The current FBI National DNA Index System (NDIS) database has near 18 million profiles (1). Shifting from sizing short tandem repeat (STR) alleles to high throughput sequencing (HTS) will enable resolution of different STR alleles with the same lengths. Illumina recently introduced the ForenSeq (2) sequencing panel that includes SNPs in addition to STRs. This expanded panel of STR and SNP profiles provides additional sample profile information for inferring biogeographic ancestry (BGA) and phenotype (externally visible traits – EVTs). Others have pioneered using SNPs for BGA prediction (3, 4), EVTs (5), mixture analysis (6), and more. Future panels will scale to thousands and tens of thousands of loci with loci selected for multiple forensics insights (6).

Previous work has examined the use of specific SNP panels for identifying populations of individuals (7, 8). This work relies on a specific assay for identifying the ancestry of an individual. This creates a limitation in the ability to apply this work to forensics applications, where SNP markers may not have been genotyped, and would impose a burden of having to re-genotype the DNA with a new panel as the assays are revised. More recently, researchers have been looking to create more generalizable algorithms that can use SNP data to predict biogeographic ancestry and descent (9-15). In this work, a generalizable framework for identifying differences between ethnicities and family members based on a subset of the available SNP data is presented.

Methods

Database Generation

SNPs were selected from the Allele Frequency Database (ALFRED) (17), containing four ethnicities with large numbers of well-characterized SNPs: African Americans, Estonians, Koreans, and Palestinians. A total of 39,108 SNPs were characterized across each of these ethnic groups. This information was used to create four initial *in silico* populations of pure ancestry, each containing 15,000 individuals. When multiple sources were listed for a single locus, the median minor allele frequency was assigned to the SNP. The gender of each individual was assigned such that

$$P_x(\text{male}) = 0.5; \quad P_x(\text{female}) = 0.5 \quad (1)$$

the probability that individual x is male is 0.5, and the probability that individual x is female is also 0.5. Minor allele information was assigned using the allele frequencies provided by ALFRED. Eight more generations were created using these four ethnic populations. Additionally, four populations of the same ethnic groups were created with a different seeding of the random number generator. This results in a starting generation of individuals that will have relatives, and an independent set of individuals with no relatives.

Within each subsequent generation a constrained pairing of a male and female from the previous generation produced offspring. The pairings generated between individuals were constrained based on previous US census information (17). Within each generation, the probability that an individual would marry given they are male is:

$$P(\text{marry } k \text{ times} \mid \text{male}) = \begin{cases} 0.336, & \text{if } k = 0 \\ 0.499, & \text{if } k = 1 \\ 0.130, & \text{if } k = 2 \\ 0.035, & \text{if } k = 3 \end{cases} \quad (2)$$

and the probability that an individual would marry given they are female is:

$$P(\text{marry } k \text{ times} \mid \text{female}) = \begin{cases} 0.278, & \text{if } k = 0 \\ 0.545, & \text{if } k = 1 \\ 0.140, & \text{if } k = 2 \\ 0.037, & \text{if } k = 3 \end{cases} \quad (3)$$

This information provides the probability that a person would get married, zero, one, two, or three or more times.

Furthermore, an ethnic intermarriage rate

$$P(\text{intermarry}) = 0.15 \quad (4)$$

was set based on a poll conducted by pew social trends (18). If an individual were the product of intermarriage, they would be given a uniform probability of marrying all individuals of the opposite gender, regardless of their ethnicity. All couples that married produced four offspring. Each child was created independently with their two alleles drawn randomly from their parents at each locus. This procedure was followed for the remaining eight generations. This in silico profiles dataset is available (16).

Characterizing Differences Between Ethnic Groups

The Euclidean distance was computed between minor allele frequencies across ethnic groups in order to better understand the distance between ethnic groups.

This distance was represented as:

$$Dist_{A_{vs}B} = \sqrt{\sum_{i=1}^{N_{SNPs}} (A_i - B_i)^2} \quad (5)$$

where A_i is the minor allele frequency for locus i within ethnicity A, and B_i is the minor allele frequency for locus i within ethnicity B.

Additionally, the fixation index (F_{ST}) was computed as a standard method for identifying population differentiation due to genetic structure. F_{ST} was defined as the difference between the total population heterozygosity (H_T) and the average subpopulation heterozygosity (H_S), all divided by the total population heterozygosity. The formula for F_{ST} is shown in equation 6.

$$F_{ST} = \frac{(H_T - H_S)}{H_T} \quad (6)$$

Characterizing Differences Between Individuals

The differences between populations were also characterized using the mean number of mismatches between individuals. Across individuals, the number of mismatches is defined by the number of occurrences where an individual has at least one minor allele at a given locus, while this allele is absent in the second individual, or when the first individual is missing a minor allele that the second individual has at least one copy of. The case where one person has a single minor allele and the second person is homozygous for the minor allele is not counted as a mismatch. This is done to allow efficient querying of minor allele differences by making a binary decision based on whether or not there exists a minor allele at the given loci. All results are reported using two subsets of the panel of SNPs. The first

subset contains 3,453 SNPs with a MAF between 0.01 and 0.1, while the second subset contains the 12,456 SNPs with a MAF between 0.01 and 0.2.

The predicted number of allele mismatches is derived using the Hardy-Weinberg equation:

$$p^2 + 2pq + q^2 = 1 \quad (7)$$

where p is the major allele frequency, and q is the minor allele frequency. Using this equation, the mean number of allele mismatches between unrelated individuals, defined here as the Combined Search Method, can be predicted as:

$$\sum_{i=1}^{N_{SNPs}} (1 - p_{Ai}^2) * p_{Bi}^2 + (1 - p_{Bi}^2) * p_{Ai}^2 \quad (8)$$

where p_{Ai} is the major allele frequency for population A, and p_{Bi} is the major allele frequency for population B, and $2pq + q^2 = 1 - p^2$. This equation provides the sum of the probabilities that an individual from population A will have at least one minor allele, while an individual from population B will have two major alleles and vice versa. Furthermore, the number of differences at the loci where an individual has a minor allele and the second individual has only the major allele, defined here as the Individual Search Method, can be predicted as

$$\sum_{i=1}^{N_{SNPs}} (1 - p_{Ai}^2) * p_{Bi}^2 \quad (9)$$

This equation provides a means for predicting the number of minor alleles that are present in an individual but absent from the population that they are being compared against.

Admixtures and Kinship Mismatches

The framework for identifying minor allele mismatches can be extended to related individuals of both pure and mixed ancestries (admixture). A person with admixed ancestry has a combination of SNP frequencies drawn proportionally from minor allele frequencies of their inherited ancestries. The admixture SNP frequencies are represented by the following formulae:

$$\sum_{j=1}^{N_{eth}} w_j * (p_{ji}^2 + 2p_{ji}q_{ji} + q_{ji}^2) \quad \forall i \in N_{SNPs} \quad (10)$$

$$\sum_{j=1}^{N_{eth}} w_j = 1 \quad (11)$$

This allows for an individual to be compared against their descendants who have a mixed ethnicity. The prediction for expected number of minor allele mismatches once again builds on Hardy-Weinberg population genetics and probability theory. Equations were derived to predict the expected number of differences between parent and child, sibling and sibling, and grandparent and grandchild.

Given a child who is the product of a parent of ethnicity A, and a parent of ethnicity B, the expected number of differences between the child and the parent of ethnicity A can be computed in two stages. The first stage is to compute the probability that a child will have a number of minor alleles $n \in \{0, 1, 2\}$, given that parent A has a defined number of minor alleles $m \in \{0, 1, 2\}$.

$$p(n | m) = p(\text{parent } B \text{ passes on } (m - n) \text{ minor alleles}) \quad (12)$$

The second stage is to multiply by the prior probability that parent A will have a number of minor alleles, m .

$$p(n \cap m) = p(n | m) * p(m) \quad (13)$$

These relationships are computed and shown in Table 1. Additionally, the final joint probability is shown for the scenario where the parents have the same ethnicity (A = B)

Using Table 2, it becomes possible to calculate the expected number of minor allele mismatches, which will be represented as the sum across loci of the probability that a child has 1 minor allele when their parent has 0 minor alleles, and the probability that the child has 0 minor alleles and their parent has 1 minor allele. In the event the parent and child are of the same ethnicity, the sum of the probabilities can be expressed as the equation:

$$2 \sum_{i=1}^{N_{SNPs}} p_i^3 q_i + p_i^2 q_i^2 \quad (14)$$

However, if the child has mixed ethnicity, with one parent of ethnicity A, and another parent of ethnicity B, then the sum of the differences will be expressed as the equation:

$$\sum_{i=1}^{N_{SNPs}} (p_{Bi} q_{Bi} + q_{Bi}^2) * p_{Ai}^2 + (p_{Bi}^2 + p_{Bi} q_{Bi}) * p_{Ai} q_{Ai} \quad (15)$$

The framework for computing the expected number of minor allele mismatches also generalizes across the type and level of relationship. In order to compute the

expected number of differences between two siblings, it is essential to account for the number of minor alleles each parent has at a given locus. Parent 1 is represented as being a member of ethnicity A, and parent 2 is represented as being a member of ethnicity B. The first equation computes the conditional probability of two siblings having a number of minor alleles $n_1, n_2 \in \{0, 1, 2\}$, conditioned on all the possible allelic combinations that their parents could have $m_1, m_2 \in \{0, 1, 2\}$. The conditional probability $p(n_1, n_2 | m_1, m_2)$ is multiplied by the prior probability $p(m_1, m_2)$, and is then marginalized over for all possible values of m_1 , and m_2 .

$$p(n_1 \cap n_2) = \sum_{m_1} \sum_{m_2} p(n_1, n_2 | m_1, m_2) * p(m_1, m_2) \forall m_1, m_2 \in \{0, 1, 2\} \quad (16)$$

$$= \sum_{m_1} \sum_{m_2} p(n_1, n_2 | m_1, m_2) * p(m_1) * p(m_2) \forall m_1, m_2 \in \{0, 1, 2\}$$

Because m_1 , and m_2 are independent events, the joint probability $p(m_1, m_2)$, can be represent as the product of $p(m_1)$ and $p(m_2)$. By marginalizing over the prior, the equation represents the full joint probability of differences between siblings. When examining differences between zero minor alleles and one or two minor alleles, the set of equations can be summarized as (zero probability parent child combinations not shown):

Using Table 2, two siblings of the same ethnicity will be represented by the equation:

$$2 * \sum_{i=1}^{N_{SNPs}} p_i^3 q_i + \frac{3}{4} p_i^2 q_i^2 \quad (17)$$

However, if the siblings have mixed ancestry, with one parent of ethnicity A, and the other of ethnicity B, the expected number of differences between siblings can be represented as:

$$\sum_{i=1}^{N_{SNPs}} \frac{1}{2} p_{Ai}^2 * p_{Bi} q_{Bi} + \frac{1}{2} p_{Bi}^2 * p_{Ai} q_{Ai} + \frac{3}{4} p_{Ai} q_{Ai} * p_{Bi} q_{Bi} \quad (18)$$

Finally, this methodology was employed to predict the expected number of minor allele mismatches between a grandparent and grandchild. This equation is derived using the chain rule of probability, alongside the conditional and prior probabilities computed in Table 1. Using the chain rule, the conditional probability is computed such that a child will have a number of minor alleles $n \in \{0, 1, 2\}$, given their grandparent has a number of minor alleles $o \in \{0, 1, 2\}$, which factors into the sum of the product of conditional probabilities of a child having n minor alleles given their parent has $m \in \{0, 1, 2\}$ minor alleles, and a parent having m minor alleles given their grandparent has o minor alleles. The summation of the conditional probabilities, multiplied by their priors is then used to compute the joint probability of a child having n minor alleles and their grandparent having m minor alleles. This results in four equations to compute the expected number of mismatches between a child and their grandparent.

$$\begin{aligned} a.) \quad p(n = 0 \cap o = 1) &= (p(n = 0 | m = 0) * p(m = 0 | o = 1) + \\ &\quad p(n = 0 | m = 1) * p(m = 1 | o = 1) + \\ &\quad p(n = 0 | m = 2) * p(m = 2 | o = 1)) * p(o = 1) \end{aligned} \quad (19)$$

$$\begin{aligned} b.) \ p(n = 0 \cap o = 2) &= (p(n = 0 | m = 0) * p(m = 0 | o = 2) + \\ &\quad p(n = 0 | m = 1) * p(m = 1 | o = 2) + \\ &\quad p(n = 0 | m = 2) * p(m = 2 | o = 2)) * p(o = 2) \end{aligned}$$

$$\begin{aligned} c.) \ p(n = 1 \cap o = 0) &= (p(n = 1 | m = 0) * p(m = 0 | o = 0) + \\ &\quad p(n = 1 | m = 1) * p(m = 1 | o = 0) + \\ &\quad p(n = 1 | m = 2) * p(m = 2 | o = 0)) * p(o = 0) \end{aligned}$$

$$\begin{aligned} d.) \ p(n = 2 \cap o = 0) &= (p(n = 2 | m = 0) * p(m = 0 | o = 0) + \\ &\quad p(n = 2 | m = 1) * p(m = 1 | o = 0) + \\ &\quad p(n = 2 | m = 2) * p(m = 2 | o = 0)) * p(o = 0) \end{aligned}$$

The summation of equations 19a through 19d over all SNPs produces the expected number of mismatches between a grandparent and grandchild.

Kinship Equations for Computing Total Differences

The framework shown above can also be used to predict total mismatches between related individuals, where a mismatch is defined as a difference between genotypes. This differs from the above methodology, as it would add in a factor to represent mismatches where an individual is heterozygous for the minor allele, and their relative is homozygous for the minor allele. Using this definition, along with Table 1,

the total expected differences between a parent and child of the same ethnicity would be:

$$2 \sum_{i=1}^{N_{SNPs}} p_i^3 q_i + 2p_i^2 q_i^2 + p_i q_i^3 \quad (20)$$

Similarly, the expected number of differences between siblings can be extended to account for the total number of differences between genotypes at the examined loci. Doing this requires the incorporation of the probability that an individual will have one minor allele, and their sibling will have two minor alleles. These probabilities are computed in Table 3.

If the siblings have two parents of the same ethnicity, then the expected number of total minor allele difference would be expressed as:

$$2 \sum_{i=1}^{N_{SNPs}} p_i^3 q_i + \frac{5}{4} p_i^2 q_i^2 + p_i q_i^3 \quad (21)$$

Finally, to predict the total expected number of differences between a grandparent and grandchild, two additional equations are incorporated.

$$\begin{aligned} e.) \ p(n = 2 \cap o = 1) &= (p(n = 2 | m = 0) * p(m = 0 | o = 1) + \\ &\quad p(n = 2 | m = 1) * p(m = 1 | o = 1) + \\ &\quad p(n = 2 | m = 2) * p(m = 2 | o = 1)) * p(o = 1) \end{aligned} \quad (19)$$

$$\begin{aligned} f.) \ p(n = 1 \cap o = 2) &= (p(n = 1 | m = 0) * p(m = 0 | o = 2) + \\ &\quad p(n = 1 | m = 1) * p(m = 1 | o = 2) + \end{aligned}$$

$$p(n = 1 | m = 2) * p(m = 2 | o = 2)) * p(o = 2)$$

Factoring equations 19e and 19f into the summation over SNPs results in the prediction of the expected number of different genotypes across all loci.

Results

Table 4 provides a summary of the Euclidean distance between each of the four ethnic groups. The Euclidean distance is computed to measure the similarity between each ethnic group as a function of the minor allele frequencies averaged over the population.

While the Euclidean distance is informative when characterizing population differences, it is less effective when comparing differences between individuals. To meet this need, changes between ethnicities and individuals across multiple generations of the *in silico* database are tracked using the number of minor allele differences. This information is then plotted as a set of kernel density estimates (KDEs). These plots provide a distribution of the number of minor allele mismatches within each comparison. Figure 1 shows a set of KDE plots with the subset of loci with a MAF between 0.01 and 0.1 (Figure 1a) and the subset of loci with a MAF between 0.01 and 0.2 (Figure 1b). These minor allele frequency sets were selection for their potential use in DNA mixture analysis. The population of 15,000 individuals is compared to the set of unrelated individuals. The number of minor allele mismatches between groups is averaged across these comparisons.

Different methodologies are compared for identifying differences between individuals and ethnic groups. Figure 2 shows two different algorithms for identifying differences. The first algorithm, shown in Figure 2a, identifies the number of differences where an individual has a minor allele at a locus, while the person they are being compared to does not. The second algorithm, shown in Figure 2b, shows the total number of allelic differences.

Figure 2 shows an interesting phenomenon. Upon utilizing the set of all minor allele differences, as shown in Figure 2b, there is no longer a guarantee that the population of individuals will be most closely related to the ethnicity with the most similar profile of minor allele frequencies. For example, while Estonians and Palestinians are shown to be the most closely related ethnic groups (Table 4), in the Estonian and Palestinian quadrants of Figure 2b, we see that the Korean population is more closely related than the most similar ethnic group. This can be explained using the information shown in Figure 2c, which plots the sorted MAF across loci. From this plot it can be observed that the Korean population has approximately 1,200 loci with a MAF approaching zero. As a result of this, when all minor allele mismatches are tracked, as is the case in Figure 2b, the Korean population will show a smaller number of mismatches than more closely related foreign populations.

Figure 3 plots the regressed best-fit lines between Euclidean distances and population distances, showing high correlation between the two methods. In

addition, a bootstrapping procedure is used to identify the 95% confidence intervals, which are shown as the shaded region around the blue curve.

The minimum number of minor allele mismatches across a population is also recorded as a way of assessing the lower limit of relation between differing populations. Figure 4 presents the minimum number of differences when comparing minor alleles found in the individual using the subset of loci with a MAF between 0.01 and 0.1, and the subset with a MAF between 0.01 and 0.2.

The *in silico* dataset allows for tracking changes in minor allele distribution across related individuals. This is demonstrated through a family that spans five generations. In each generation, some of the children married within their own ethnic background, while others married outside of their ethnic group. The observed intermarriage results in descendants with a varying percentage of their original ethnicity. This is demonstrated in Figure 5, which tracks an African American family across four generations. Each quadrant represents a degree of difference from the original ancestor (i.e. 1° = children, 2° = grandchildren, 3° = great grandchildren, 4° = great great grandchildren).

As can be observed from Figure 5, a greater dilution in the original ethnicity results in a greater number of mismatches across generations. This will result in someone with 75% of the original ethnicity with a second degree relationship, appearing similar to someone with 100% of the parent ethnicity and a third degree

relationship. In order to capture this affect across families, mismatches between thousands of descendants was tracked at the parent-child, and grandparent-grandchild level. Figure 6 shows the average level of parent-child, and grandparent-grandchild relation, when the descendant has 100%, 75%, 50%, and 25% of the ethnicity of the progenitor. Each descendant has at most two races that go into their ethnic background.

Discussion

The probability equations used to estimate the number of mismatches between related and unrelated individuals builds on the work of Cassa *et al* (18). However, while their work looks to identify how an individual's genotype can be inferred from their sibling's genotype, this work builds out the relationship between individuals of different levels of relationship, and then compares their expected values. This paper provides an in-depth analysis of changes in minor allele representation across separate ethnic groups as well as through generations of genetic inheritance.

Table 4 provides a summary of the Euclidean distance between each of the four ethnic groups. This provides a theoretical basis for interpreting the *in silico* results. Each ethnic group will be closest to their own ethnicity, followed by each of the remaining groups. Based on Table 4, African Americans are predicted to be most similar to Palestinians, followed by Estonians, and finally Koreans. Estonians will have the most in common with Palestinians, then Koreans, and finally African Americans. Koreans are closest to Estonians, followed by Palestinians, and then African Americans. Finally, Palestinians are closest to Estonians, followed by African

Americans, and finally Koreans. The Euclidean distances also show that the shortest allelic distance between differing ethnicities is between Estonians and Palestinians. This information is concordant with the F_{ST} numbers that have also been shown in Table 4. Additionally, Figure 3 demonstrates that there is a strong correlation with narrow error bars between the Euclidean distances and the population distances. This creates a method for identifying similarity between ethnic groups, which then provides information on the expected distance between individuals in two different populations.

The curves shown in Figure 1a and Figure 1b demonstrate the increased ability to separate populations of individuals as a function of the number of incorporated SNPs. Figure 1b shows that the increased subset of SNPs provides a greater degree of separation between ethnic groups. Upon examining this plot, it can be observed that the order of closest ethnicity matches what is shown in Table 4 by Euclidean distances. The only exception to this is shown in the panel of Estonian plots where the African American curve is closer than the Korean curve. However, in this case the African American and Korean curves are significantly overlapping. Figure 2a and Figure 2b show a comparison of the Individual and Combined methods for identifying minor allele differences between individuals. The combined method provides for symmetry of comparisons, while the individual methods tailors the comparison to the minor alleles present in the person of interest. As shown in the combined method, this property of symmetry can cause a group of unrelated individuals from a particular ethnic group to have a distribution that is consistently

closer to the individual's ethnicity. This is a function of the number of loci with a minor allele frequency of approximately zero, as shown in Figure 2c. This large number of alleles with a near zero MAF cause a greater prevalence of the major allele, leading to less allelic mismatches with the combined method. This result causes the Korean population to be consistently closer in distribution to the other ethnic groups.

The comparative methods are extended to kinship generations with admixture. As shown in Figure 6, these methods can be employed to identify related individuals of mixed race. However, as the fraction of the progenitor's original ancestry becomes minimized, it becomes more difficult to separate out related individuals from unrelated individuals of the same ethnicity.

Conclusions

This paper has presented a framework for characterizing identity search results, which has been used to compare both related and unrelated individuals from four different ethnicities. In the process of distinguishing populations of different ethnicities, this work has shown that ethnicity impacts the expected number of minor allele mismatches as expected. Furthermore, this method has been applied to related individuals, where it has been shown that relatives with admixture ancestry have more minor allele mismatches than relatives with shared ancestry.

Acknowledgements

Distribution Statement A. Approved for public release: distribution unlimited.

This material is sponsored by the Defense Threat Reduction Agency under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

Table 1: Table of probability equations used to estimate the number of mismatches between a child and their parent. The child is compared to a parent of ethnicity A, and has another parent of ethnicity B. The joint probability is shown for the condition when the parents are of the same ethnicity.

<u>Child MA</u>	<u>Parent MA</u>	<u>Cond. Prob.</u>	<u>Prior Prob.</u>	<u>Joint Prob. (A = B)</u>
0	0	$p_B^2 + p_B q_B$	p_A^2	$p^4 + p^3 q$
1	0	$p_B q_B + q_B^2$	p_A^2	$p^3 q + p^2 q^2$
2	0	0	p_A^2	0
0	1	$\frac{1}{2}(p_B^2 + p_B q_B)$	$2p_A q_A$	$p^3 q + p^2 q^2$
1	1	$\frac{1}{2}(p_B^2 + p_B q_B) + \frac{1}{2}(p_B q_B + q_B^2)$	$2p_A q_A$	$p^3 q + 2p^2 q^2 + p q^3$
2	1	$\frac{1}{2}(p_B q_B + q_B^2)$	$2p_A q_A$	$p^2 q^2 + p q^3$
0	2	0	q_A^2	0
1	2	$p_B^2 + p_B q_B$	q_A^2	$p^2 q^2 + p q^3$
2	2	$p_B q_B + q_B^2$	q_A^2	$p q^3 + q^4$

Table 2: Table of probability equations used to estimate the number of mismatches between an individual and their siblings. The siblings have two parents, one of ethnicity A, and the other of ethnicity B. The joint probability is shown for the condition when the parents are of the same ethnicity.

<u>Sib1, Sib2</u>	<u>Parent1,</u> <u>Parent2</u>	<u>Cond. Prob.</u>	<u>Prior Prob.</u>	<u>Joint Prob. (A = B)</u>
0, 1	0, 1	$\frac{1}{4}$	$p_A^2 * 2p_Bq_B$	$\frac{1}{2}p^3q$
0, 1	1, 0	$\frac{1}{4}$	$p_B^2 * 2p_Aq_A$	$\frac{1}{2}p^3q$
0, 1	1, 1	$\frac{2}{16}$	$2p_Aq_A * 2p_Bq_B$	$\frac{1}{2}p^2q^2$
0, 2	1, 1	$\frac{1}{16}$	$2p_Aq_A * 2p_Bq_B$	$\frac{1}{4}p^2q^2$

Table 3: Table of probability equations used to estimate the number of genotype mismatches between an individual and their siblings. The siblings have two parents, one of ethnicity A, and the other of ethnicity B. The joint probability is shown for the condition when the parents are of the same ethnicity. This table extends upon Table 2.

<u>Sib1, Sib2</u>	<u>Parent1,</u> <u>Parent2</u>	<u>Cond. Prob.</u>	<u>Prior Prob.</u>	<u>Joint Prob. (A = B)</u>
1, 2	1, 1	$\frac{2}{16}$	$2p_Aq_A * 2p_Bq_B$	$\frac{1}{2}p^2q^2$
1, 2	1, 2	$\frac{1}{4}$	$2p_Aq_A * q_B^2$	$\frac{1}{2}pq^3$
1, 2	2, 1	$\frac{1}{4}$	$2p_Bq_B * q_A^2$	$\frac{1}{2}pq^3$

Table 4: Population distance metrics between each of the four ethnic groups

- a) Euclidean distance for the 3,453 SNPs with a MAF between 0.01 and 0.1.
- b) Euclidean distance for the 12,456 SNPs with a MAF between 0.01 and 0.2.
- c) F_{ST} for the 3,453 SNPs with a MAF between 0.01 and 0.1.
- d) F_{ST} for the 12,456 SNPs with a MAF between 0.01 and 0.2.

(4a)

Ethnicity	African Am.	Estonian	Korean	Palestinian
African Am.	0	5.48	5.62	5.17
Estonian	5.48	0	5.27	3.56
Korean	5.62	5.27	0	5.34
Palestinian	5.17	3.56	5.34	0

(4b)

Ethnicity	African Am.	Estonian	Korean	Palestinian
African Am.	0	16.96	19.73	15.90
Estonian	16.96	0	15.48	8.52
Korean	19.73	15.48	0	16.05
Palestinian	15.90	8.52	16.05	0

(4c)

Ethnicity	African Am.	Estonian	Korean	Palestinian
African Am.	0	0.0299	0.0358	0.0274

Estonian	0.0299	0	0.0302	0.0114
Korean	0.0358	0.0302	0	0.0311
Palestinian	0.0274	0.0114	0.0311	0

(4d)

Ethnicity	African Am.	Estonian	Korean	Palestinian
African Am.	0	0.0458	0.0620	0.0410
Estonian	0.0458	0	0.0404	0.0122
Korean	0.0620	0.0404	0	0.0434
Palestinian	0.0410	0.0122	0.0434	0

Table 5: The population distance between each of the four ethnic groups

- a) Shown for the 3,453 SNPs with a MAF between 0.01 and 0.1 using the individual's minor alleles.
- b) Shown for the 12,456 SNPs with a MAF between 0.01 and 0.2 using the individual's minor alleles.
- c) Shown for the 12,456 SNPs with a MAF between 0.01 and 0.2 using all allelic differences.

(5a)

Ethnicity	African Am.	Estonian	Korean	Palestinian
African Am.	284.25	313.88	343.41	316.36
Estonian	527.54	458.82	531.69	486.86
Korean	289.40	264.01	244.61	273.22
Palestinian	468.27	425.12	479.12	412.63

(5b)

Ethnicity	African Am.	Estonian	Korean	Palestinian
African Am.	1675.26	2092.78	2229.77	2054.40
Estonian	2510.69	2149.04	2483.20	2259.96
Korean	2110.76	1946.21	1624.91	1983.25
Palestinian	2409.03	2196.67	2456.84	2105.51

(5c)

Ethnicity	African Am.	Estonian	Korean	Palestinian
African Am.	3350.53	4603.64	4340.68	4463.54
Estonian	4603.47	4289.23	4429.61	4456.74
Korean	44340.57	4429.52	3429.99	4440.27
Palestinian	4463.45	4456.79	4440.21	4211.14

References

1. CODIS - NDIS statistics. [cited]; Available from: <https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics>.
2. Hussing C, Børsting C, Mogensen HS, Morling N. Testing of the Illumina(R) ForenSeq™ kit. *Forensic Sci Int Genet Suppl*. 2015;5:e449-e50.
3. Keating B, Bansal AT, Walsh S, Millman J, Newman J, Kidd K, et al. First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. *Int J Legal Med*. 2013;127(3):559-72.
4. Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investigative Genet*. 2013;4(1):13.
5. Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, et al. The HirisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet*. 2013;7(1):98-115.
6. Ricke D, Shcherbina A, Chiu N, Schwoebel E, Harper J, Petrovick M, et al., editors. Sherlock's Toolkit: A forensic DNA analysis system. 2015 IEEE International Symposium on Technologies for Homeland Security (HST); 2015 Apr 14-16.
7. Rogalla U, Rychlicka E, Derenko MV, Malyarchuk BA, Grzybowski T. Simple and cost-effective 14-loci SNP assay designed for differentiation of European, East Asian and African samples. *Forensic Sci Int Genet*. 2015;14:42-9.

8. Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B.
Selection of highly informative SNP markers for population affiliation of major US
populations. *Int J Legal Med.* 2016;130(2):341-52.
9. Zheng X, Weir BS. Eigenanalysis of SNP data with an identity by descent
interpretation. *Theor Popul Biol.* 2016;107:65-76.
10. Prestes PR, Mitchell RJ, Daniel R, Sanchez JJ, van Oorschot RAH. Predicting
biogeographical ancestry in admixed individuals – values and limitations of using
uniparental and autosomal markers. *Aust J Forensic Sci.* 2016;48(1):10-23.
11. Verdu P, Rosenberg NA. A general mechanistic model for admixture histories
of hybrid populations. *Genet.* 2011;189(4):1413-26.
12. Conomos MP, Miller MB, Thornton TA. Robust Inference of Population
Structure for Ancestry Prediction and Correction of Stratification in the Presence of
Relatedness. *Genet Epidemiol.* 2015;39(4):276-93.
13. Korostishevsky M, Malkin I, Spector T, Livshits G. Parametric model-based
statistics for possible genotyping errors and sample stratification in sibling-pair SNP
data. *Genet Epidemiol.* 2010;34(1):26-33.
14. Yang H-C, Wang P-L, Lin C-W, Chen C-H, Chen C-H. Integrative analysis of
single nucleotide polymorphisms and gene expression efficiently distinguishes
samples from closely related ethnic populations. *BMC Genomics.* 2012;13(1):346.
15. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free Estimation of
Recent Genetic Relatedness. *Am J Human Genet.* 2016;98(1):127-48.
16. 11 Million SNP Profiles datasets. [cited June 1, 2018]; Available from:
<https://doi.org/10.7910/DVN/IDT8HZ>.

17. Kidd KK. The ALlele FREquency Database. 2017 [updated 2017; cited];

Available from: <https://alfred.med.yale.edu/>.

18. The Rise of Intermarriage. 2012 [updated 2012; cited]; Available from:

<http://www.pewsocialtrends.org/2012/02/16/the-rise-of-intermarriage/>.

Figure Legends

FIG 1. Mean Mismatch Individual Search Results: Blue = African American, Green = Estonian, Red = Korean, Purple = Palestinian, Black = Estimates.

- a) Shows the KDE for the mean comparison against the individual's minor alleles with MAF between 0.01 and 0.1 (3,453 SNPs).
- b) Shows the KDE for the mean comparison against the individual's minor alleles with MAF between 0.01 and 0.2 (12,456 SNPs).

FIG 2. Mean Mismatch Individual and Combined Search Results: Blue = African American, Green = Estonian, Red = Korean, Purple = Palestinian. Black = Estimates.

a) Shows the KDE for the mean comparison against the individual's minor alleles.

b) Shows the KDE for the mean comparison of all combined allelic differences.

c) Shows the sorted MAF for each ethnic group.

FIG 3. The four curves represent the correlation between Euclidean distances and population distances (that include all allelic differences) across the four ethnic groups: Blue = African American, Green = Estonian, Red = Korean, Purple = Palestinian. The African American correlation includes bootstrapped 95% confidence intervals.

FIG 4. Minimum Mismatch Individual Search Results: Blue = African American, Green = Estonian, Red = Korean, Purple = Palestinian. Black = Estimates.

- a) Shows the KDE for the minimum comparison against the individual's minor alleles with MAF between 0.01 and 0.1.
- b) Shows the KDE for the minimum comparison against the individual's minor alleles with MAF between 0.01 and 0.2.

FIG 5. Results for close relatives including admixture. (Percentages identify the fraction of progenitor ethnicity shown in the distribution)

a) Shows the KDE for 12,456 SNPs with a MAF between 0.01 and 0.2.

b) Shows the KDE for 3,453 SNPs with a MAF between 0.01 and 0.1.

FIG 6. Kinship generations with admixture. (Blue = Gen2, 100% progenitor race; Green= Gen2, 50% progenitor race; Red = Gen3, 100% progenitor race; Yellow = Gen3, 75% progenitor race; Purple = Gen3, 50% progenitor race; Cyan = Gen3, 25% progenitor race; Gray = Unrelated progenitor race; Black dash = Predictions based on probability equations)